

An open source solution for document preservation to enable information generation from data for future generations of researchers

Fall 2014 AGU Meeting
San Francisco, CA Dec. 15-20

mo.khayat@nasa.gov
http://disc.gsfc.nasa.gov



Mo Khayat^{1,2}, Steve Kempler¹, Barbara Deshong^{1,2}, James Johnson^{1,2},
Irina Gerasimov^{1,2}, Ed Esfandiari^{1,2}, Michael Berganski^{1,2}, Jennifer Wei^{1,2}
¹NASA Goddard Space Flight Center, ²ADNET Systems Inc.

The case for document preservation

Many NASA Earth Observing System (EOS) missions have either reached the end of their active life or are nearing it. In order to ensure that future users can draw maximum benefit from this data for years to come, it is necessary for us not only to ensure preservation of the data proper, but also the preservation of all associated metadata, calibration/validation data, and important documentation or other artifacts. Missing any information that could hamper the Data-Information-Knowledge chain takes away from the ability of future generations of researchers to draw value from our data. Preservation of data products is a fairly well defined task for the NASA EOS Data Centers. However, supporting documentation and other artifacts from these missions are also critical to the long-term studies of our planet's climate. To be successful in this goal, we need to shift our focus from strict Data Lifecycle Management, to the lifecycle management of information and what enables knowledge into the future. A break in that chain, like having incomplete documentation, can be deleterious for legacy missions especially when original investigators or people most familiar with the data have long moved on or are no longer accessible.

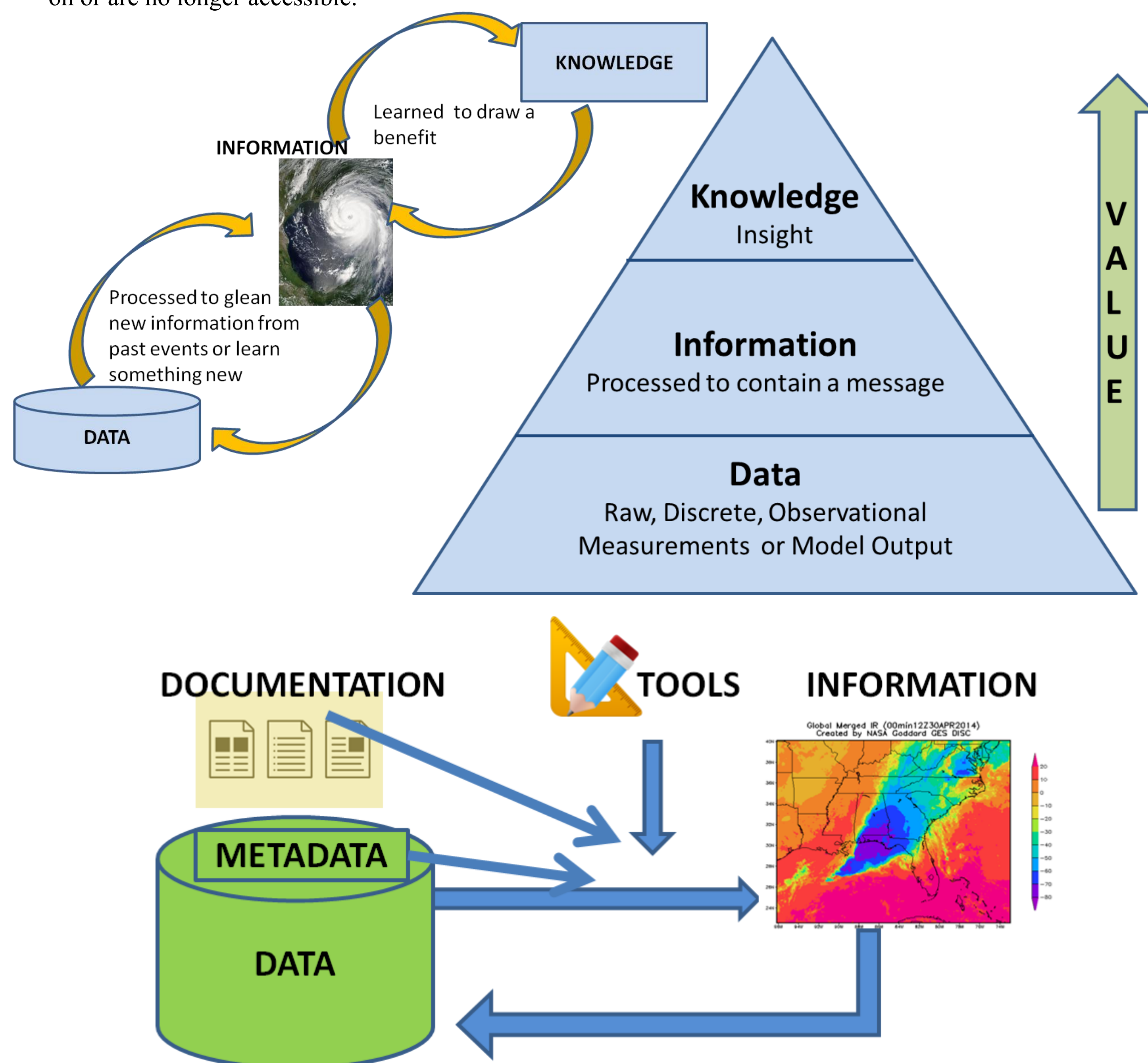


Figure 1. Data to Knowledge pathway requires information that may be contained in documents created throughout the life of a mission. Loss of such documentation could potentially deprive future generation of derived values from these data.

The Need for Documentation Preservation at NASA GES DISC

Given that NASA is not legislatively mandated to preserve data permanently, unlike agencies like USGS, NOAA and NARA, we have a challenge to develop a low cost solution that meets our data center needs and that of our users. The Goddard Earth Sciences Data and Information Services Center (GES-DISC) has implemented a Repository System to facilitate the long-term archive of documentation artifacts and other associated digital content. The GES-DISC designed this system based on Fedora Commons, an open-source repository management software, for cost savings and flexibility.

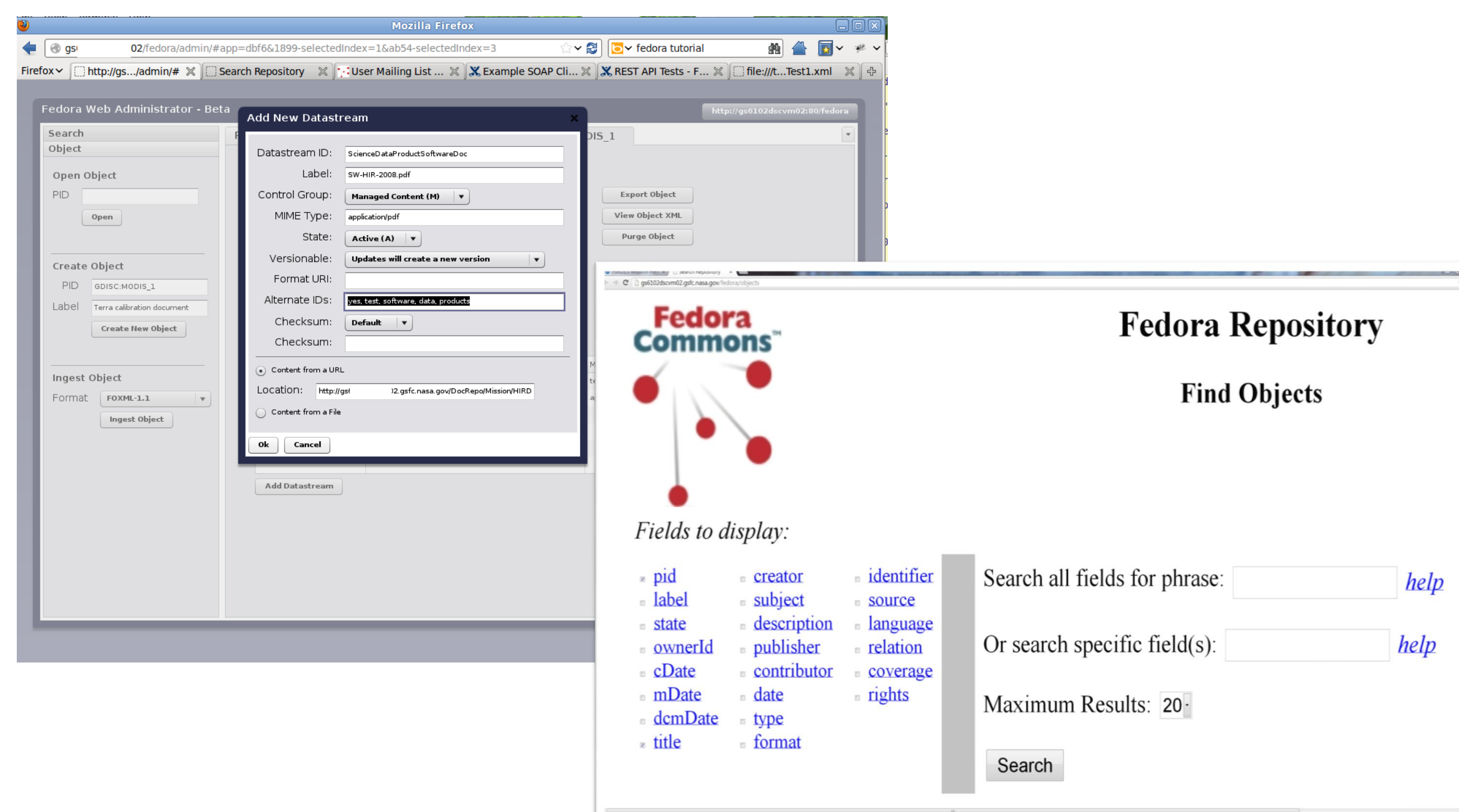
The first mission to utilize the GES-DISC Repository System was the High Resolution Dynamics Limb Sounder (HIRDLS) on the Aura spacecraft. Data and documentation from the Upper Atmosphere Research Satellite (UARS) and the Total Ozone Mapping Spectrometer (TOMS), and Nimbus have also been added. The GES-DISC is negotiating the transfer of data preservation items from the current Microwave Limb Sounder (MLS) on Aura, and the Atmospheric Infrared Sounder (AIRS) missions before they end.

Fall 2014 AGU Meeting, Poster No. IN53C-3816

Fedora Commons Interface

The GES-DISC used Fedora Commons, an open-source repository management system that is used in many universities, research centers, and libraries. It comes with a simple web-based GUI interface which provides for easy administration of the system. The GUI also allows one to enter objects or datastreams (these can be of any type document, image, source code, binary data, etc.) into the system. The system uses XML to manage the objects. The GES-DISC has also developed a command line script to allow batch ingest of objects into the Fedora Repository.

Figure 2. Internal GUI used to ingest and archive records into the repository (left), and internal GUI used to search and retrieve records from the repository (right).



Public Access of Preservation Documents

External users access the publicly available documents by visiting the mission specific documentation page for that instrument. The Fedora repository system is at the backend and makes access to the linked documents possible. Note that restricted objects (ITAR, proprietary, or software) are not accessible through the public interface. Three missions are now public:

- HIRDLS <http://disc.sci.gsfc.nasa.gov/Aura/additional/documentation/hirdls-preservation-documents>
- TOMS <http://disc.sci.gsfc.nasa.gov/acdisc/documentation/toms-mission-preservation-documents>
- UARS <http://disc.sci.gsfc.nasa.gov/acdisc/documentation/uars-mission-preservation-documents>

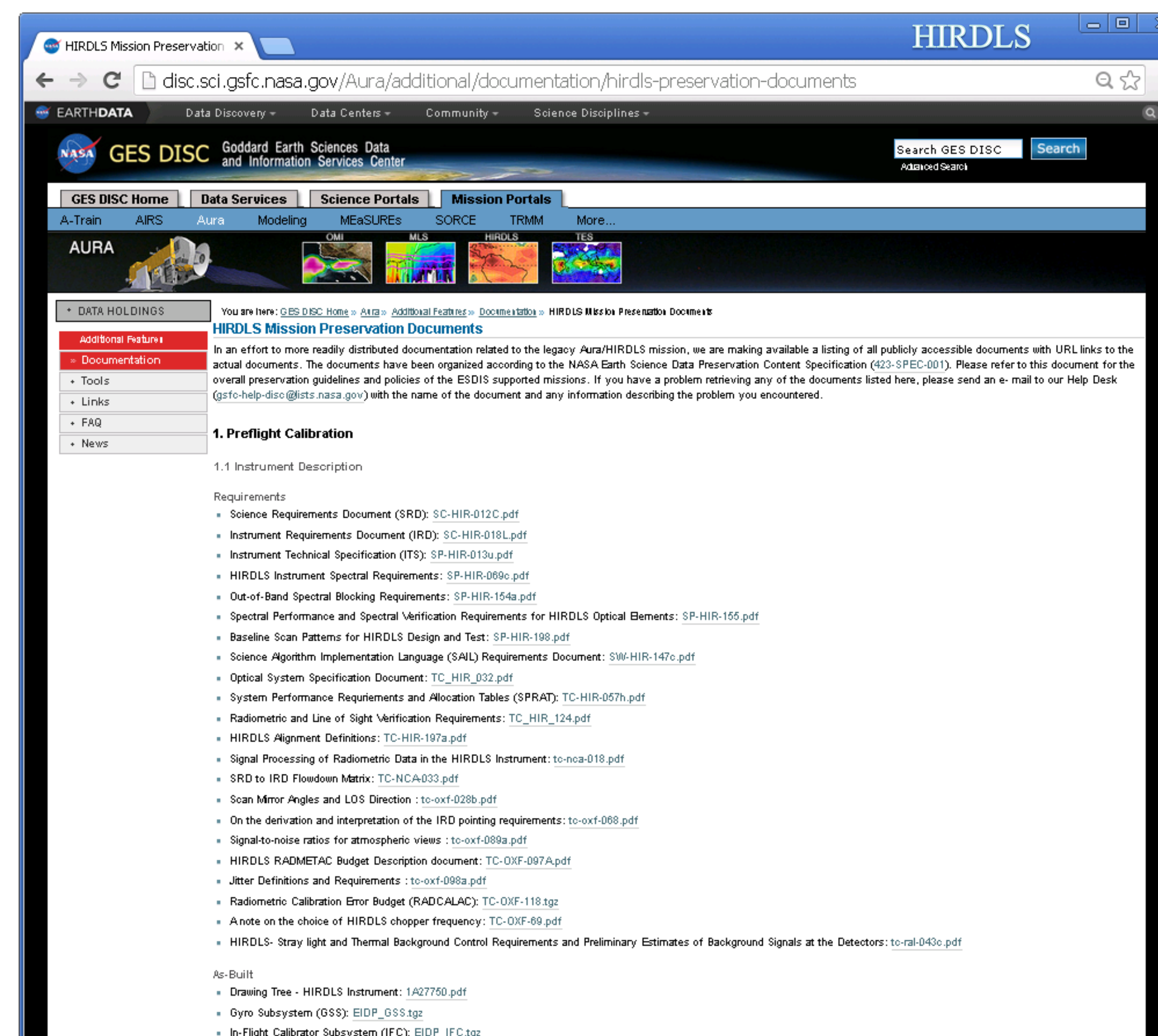


Figure 3. Example of the public web sites for the HIRDLS Mission showing preservation documents available to GES-DISC users.

GES DISC Preservation Implementation

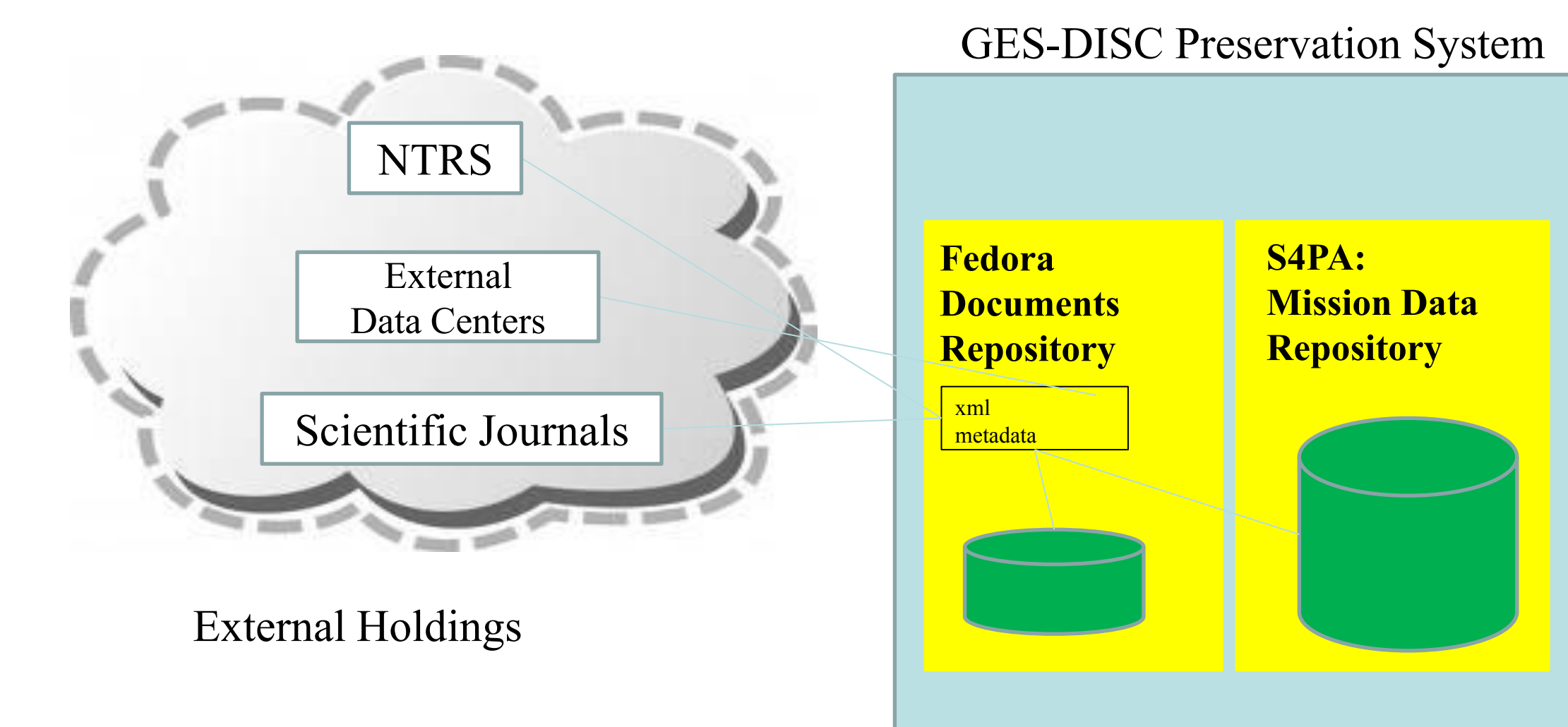
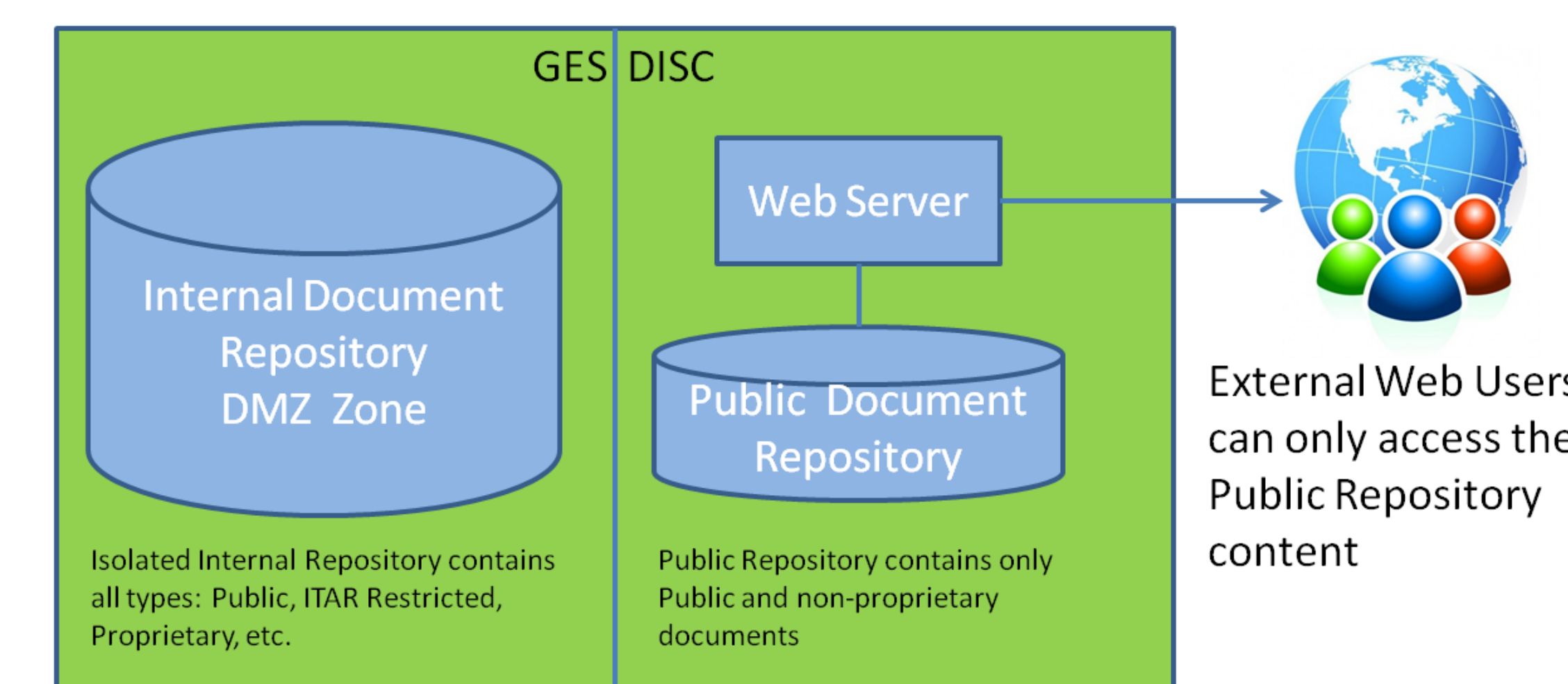
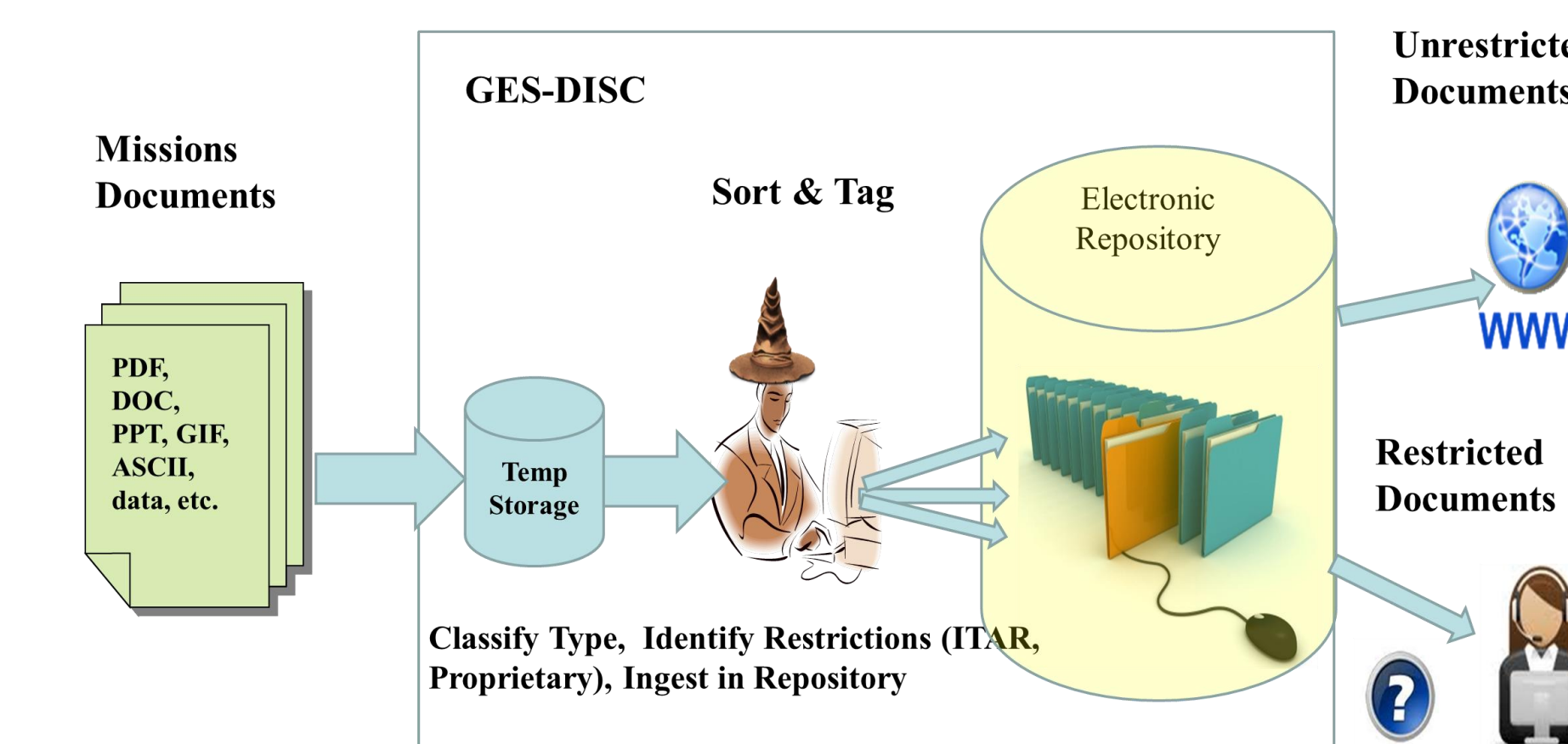


Figure 4 - Overview of the GES DISC data and documentation preservation systems. Artifacts may reside at the GES DISC or be external, (e.g. NASA Technical Reports Server (NTRS), another Data Center, or in a scientific or technical journal)

1. Identify documentation
2. Specify and implement preservation environment
3. Retrieve documentation
4. Implement retrieval and distribution services
 1. Access for internal GES-DISC users
 2. External Access via WWW for unrestricted documents
 3. External Access for restricted documents (ITAR) via User Services contact
 4. External Access for restricted documents via User Services contact

Figure 5 - Overview of the physical objects sorting, tagging, storage in archive and distribution system. (Right)

Restricted documents are isolated on an internal GES DISC accessible network only. (Below)



Lessons Learned, Challenges, and Future Plans

- Heritage missions require extensive work to identify and classify documents
- Restricted (ITAR or Proprietary) vs. Unrestricted requires special handling on a case-by-case basis (subject to export control rules)
- Limited ability to use NASA infrastructure like NTRS which are set to accept all Science and Technical Information (STI) that missions generate.
- Incorporate DOI metadata into repository (future plan)
- Upgrade from current Fedora Commons 3 to version 4 (planned release in 2015)

References

NASA Earth Science Data Preservation Content Specification (423-SPEC-001) H. K. Ramapriyan, EOSDIS Project Office, NASA GSFC https://earthdata.nasa.gov/sites/default/files/field/document/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf

Evolution of Information Management at the GSFC Earth Sciences (GES) Data and Information Services Center (DISC), IEEE Transactions on Geoscience and Remote Sensing, Volume 47, Issue: 1, 2009